Statistical Patterns of Diacritized and Undiacritized Yorùbá Texts

Asubiaro, Toluwase
E. Latunde Odeku Medical Library
College of Medicine
University of Ibadan
Nigeria
toluwaase@gmail.com

ABSTRACT: *Yorùbá standard orthography involves heavy use of diacritics for tone marking and representation of characters that are beyond ANSI scope. The diacritics are not always applied in many Yorùbá documents because specialized and language-dependent input devices for the language are very rarely available. Hence, this study aims at explicating the statistical implication of the inconsistency in the use of diacritics in electronic Yoruba documents on the distribution of word in the two versions of its texts. This was achieved by modeling the texts of Yoruba language based on Zipf's and Heap's law on the n-grams (for n=1, 2 and 3) with corporal of 1,089,318 words that are diacritically marked and its version that are unmarked diacritically. It was observed that the Zipf's graphs of the two corporal exhibited no significant difference. On the other hand, the Heap's graphs of the diacritized and undiacritized texts deviated significantly from the base. This shows that the use of the diacritics significantly affect single word distribution of the language but the effect reduced in the distribution of co-occurrences of two or more words.*

Keywords: Zipf's Law, Heaps law, Yorùbá Language, Diacritics, Statistical Language Model, Word Distribution

Received: 22 May 2015, Revised 19 June 2015, Accepted 28 June 2015

1. Introduction

Diacritics include sub-dots and tone marks which are appended to base or American Standard Institute (ANSI) characters. Diacritics are appended on base characters to represent some speech sounds that are beyond the scope of ANSI conventional codes for writing which is based on Latin encoding system. Hence, diacritics extend the functionality of these base characters, therefore new characters are formed by appending diacritic mark(s) on a base character. For instance, when a sub-dot is appended to „s character, a new character c is formed. In some languages like Yorùbá, tonality is represented with the tone marks; high tone (\) and low tone (/) which are applied on its vowels and nasal consonant. Yorùbá also cater for speech sounds that are not represented in the 26 alphabets of Latin encoding from which it inherited its writing style. These characters are ṣ, ẹ, ọ

Like Yorùbá, some African and European languages such as Hausa, Igbo, French, German, Italian and Finnish use diacritics on some base characters. While diacritics carry morphological information in some of these languages, in others, diacritics do not.

In Yorùbá, German and Finnish for instance, the use of diacritics provide morphological and lexical information. "Ojo", "ọjọ", "òjò", "òjó" for instance, are different Yorùbá words derived by appending diacritical marks on "ojo", each has a distinct meaning which differs from others derived from the same base characters. Italian and French languages also use diacritics, but the use of diacritics bear insignificant morphological or lexical information.

When texts of languages that heavily use diacritics are normalized; that is the diacritics are removed or they are not appended on necessary words, information is lost or distorted in such texts. The statistical properties of the texts may also be affected. The four variants of "ojo" which appear and are distributed accordingly as four words in properly diacritically marked Yoruba text only appear as a word –"ojo" if the texts are otherwise unmarked with diacritics. It could therefore be hypothesized that the statistical properties of the two versions of the 'orthographies' of a language could be different.

Statistical properties of written texts have been observed to follow some universal regularities. These regularities are studied in Statistical language modeling (SLM) by attempting to understand human languages through the observations of the regularities. SLM is an attempt to capture and compute a probability distribution of word or character sequences in natural languages, such that sequences which are well-formed are given a higher likelihood than those which are not [1], [2]. SLM studies have informed research work in development of language technologies. Statistical properties of written texts such as the distribution of word frequencies and increase of the vocabularies or distinct words are some of the various universal regularities observed in SLM have been modeled by Heaps law. Heaps law is a power law which explains that the number of distinct words or vocabulary of a given language will increase slowly with the increase in its document size. Accordingly, for a language with a number of collections of written texts or spoken speech and $V(n)$ estimated number of unique or distinct words in a collection n, while T is the number of tokens in the collection, the relation $V(n) = KT^{\alpha}$ holds where $0 < \alpha < 1$. Heaps Law predicts the vocabulary size of the texts of a given language from the size of a text [3].

Another law that have modeled human languages which have co-existed with Heaps law in studies is the Zipf s law. It explains that in a sample of written texts or soken speech of a given language, the few very high frequency words account for bigger proportion of the text size or spoken speech in a language. There is an approximate mathematical relation between the frequency of occurrence of each distinct word denoted as f and its rank in .The mathematical relationship between the frequency of occurrence of a given distinct word or vocabulary denoted as f and its rank r, it was given as

$$f = 1/r^{\alpha}$$

Empirically, when the list of all the words used in the text are ordered by decreasing frequency, the relationship between each distinct word and its rank as given in equation 1 is an inverse power law with an exponent that is close to 1[4], [5]

Many human languages, mostly of the indo-European origins have been found to conform to the Zipf s and Heaps law [6], [7], [8], [9], [10], [11]. These laws are co-efficients of these laws depend on language[10]. Apart from this, [12], [13] found that randomly generated texts and index terms obeyed these laws. On the other hand, it has been found that it does not hold for raw Asian languages like Chinese, Korean and Japanese [14], but holds for word segmented corpus of Chinese [15]. Zipf s and Heaps laws are both power laws which have been found to be theoretically and empirically related [16], [17], [18], [19], [20].

2. Yoruba Language and Its Orthography

Yorùbá language is spoken by over 30 million people in different parts of the world. Its native name is "*ede Yorùbá*". The native speakers of Yorùbá language occupy the southwestern part of Nigeria, a part of southern Benin Republic and southern Togo. There are traces of the use of the language in Santeria religion as language of worship where is called "*Lucumi*" or "*Nago*" in Argentina, Cuba, Puerto Rico and the Dominican Republic. There are also reported traces of the use of the language by some natives in Sierra Leone where it is called "*oku*". [21], [22].

Standard Yorùbá orthography demands a heavy use of diacritically marked characters (sub-dots and tone marks). Diacritics are used for marking tonality and to cater for the need to represent speech sounds that are beyond the range of the basic America National Standard Institute (ANSI) characters or standard Latin encoding system. It should be noted that the conventional computer keyboards is based on ANSI convention. These characters; ẹ, ẹ, ọ, ọ,á, à, è, é, í, ì, ò, ó, ù, ú, ọ, ẹ, ṣ, ń, and ǹ which are used in Yorùbá orthography and are beyond ANSI scope therefore do not appear on the conventional keyboards.

However, due to dearth of specialized and Yorùbá language-dependent device input device that could adequately and speciallycater for these diacritically marked characters, these diacritically marked characters are mostly represented electronically with the available equivalent ANSI character which are the equivalent ANSI diacritically unmarked characters. The base characters of the diacritically marked characters are also their unmarked equivalents. For instance, characters "ẹ, ẹ, è, é and ẹ" are all represented by their unmarked equivalent; "e". These practices are either partial where the diacritics are correctly applied on choice words or total.

In a previous study[10], it was proved that SLM like Heaps laws are language dependent. In essence, this study proposes a hypothesis that Heaps behavior of a language is orthographically dependent. There are two versions of the orthography of Yoruba language: the standard and the sub-standard. The standard orthography of Yoruba requires heavy use of diacritics for tone marking and representation of characters that are beyond the ANSI characters. While the sub-standard version of the orthography does not append the diacritics (in other words, characters with diacritics are normalized). Most computer encoded Yoruba texts fall to the sub-standard orthography category.

3. Methodology

The word list of n-grams (unigram, bigram and trigram) was obtained for the two corpora and ranked in decreasing order of frequency of occurrence. For Zipf s graph, logarithmic values of frequency (*Fr*) were plotted against logarithmic value of rank (*r*). For Heaps graph, *V(n)* was estimated as the number of distinct or unique words in each collection, while *T* is the number of tokens in the collection. For Heaps graph, values of *V(n)* was plotted against the values of *T*.

Text corpus that is representative, orthographically accurate and large enough is very essential in linguistics and language processing studies. Yorùbá language lacks corpus for linguistic experiments. The first step taken was gathering data set that could be acceptable in quantity and quality for the study. Texts were collected online and offline. The sources of data collected is displayed on Table 1.12% of the texts used for this study were news articles collected online.

This is consistent with TREC s methodology of using news articles for corpus development. A corpus of 1,089,318 was used for the study.

To obtain diacritically marked version of texts that were originally not appended with diacritics, they were automatically diacritized. The diacritics were also removed from the originally diacritically marked texts to obtain its diacritically unmarked version. In this paper, the diacritically marked and unmarked texts are referred to as the diacritized and undiacritized texts respectively.

	Source	**No of Articles**	**Corpus Size**
Originally undiacritized online)	Alaroye (Yorùbá weekly newspaper published	782	676,634
Originally diacritized	Yorùbá Published novels (collected offline)	4	165,553
Originally undiacritized	Academic Projects written in Yorùbá language (collected online)	10	203,416
Originally undiacritized	Yorùbá Online (Yorùbá online news collected	49	43,715
		Total	1,089,318

4. Results and Discussion

Table 2 shows rank- distribution of the ten most frequent words in the diacritized and undiacritized Yoruba texts. The table explains the word-frequency of Yoruba texts as they are affected by the use or non-use of diacritics.

Table 2. Word frequency of the diacritized and undiacritized Yorùbá most frequent unigrams

	Diacritized Texts		Undiacritized Texts	
Rank	**Index Term**	**Frequency**	**Index Term**	**Frequency**
1	tí	35818	ti	45063
2	ni	34794	ni	42659
3	wọ́n	27353	won	33258
4	àwọn	24913	o	25043
5	D	21904	awon	24953
6	ó	21028	si	23903
7	pé	20349	n	22439
8	tó	19748	pe	21310
9	kò	19167	ko	21222
10	náà	16736	to	19947

Zipf's Law

The Zipf's graphs of unigram, bigrams and trigrams of diacritized and undiacritized Yoruba texts are presented on Figures 1a, 1b and 1c respectively. The three graphs show that the diacrititized and undiacritized texts converged on most regions of the Zipf 's graphs. This shows that the diacritized and undiacrized Yoruba texts on Zipf's graph are not significantly different. This is further proved with the R^2 value of the straight line graph drawn on the Zipf's curve. The R^2 value for unigram of the diacritized and undiacritized are 0.98 and 0.97 respectively. For the bigrams, R^2 for diacritized and undiacritized are 0.95 and 0.94 respectively, while for the trigram, R^2 for diacritized and undiacritized are 0.84 and 0.85 respectively.

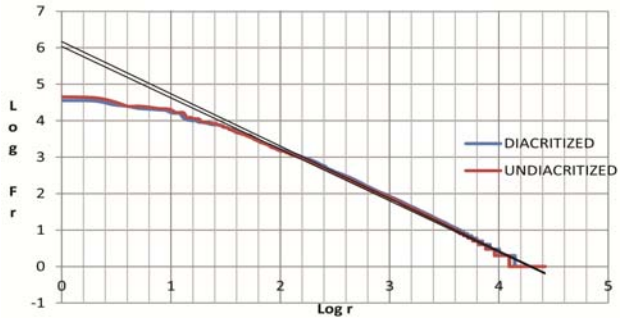

Figure 1a. Zipf's Graph for Unigram

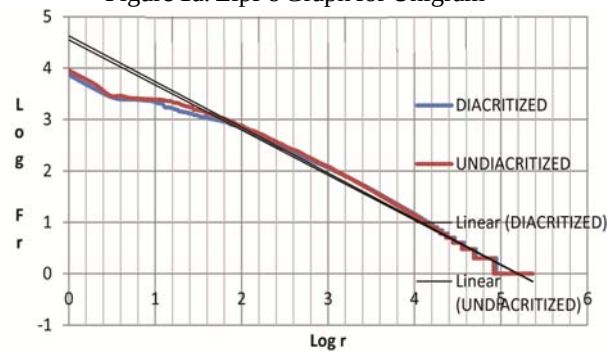

Figure 1b. Zipf's Graph for Bigram

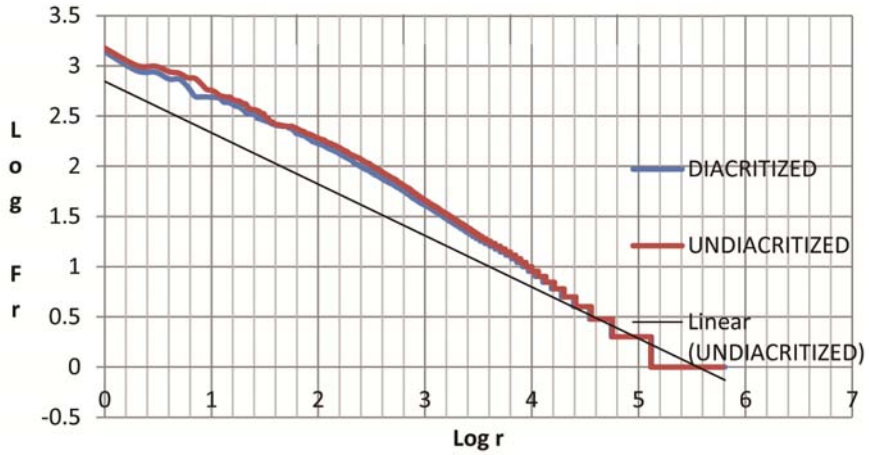

Figure 1b. Zipf's Graph for Trigram

Heaps' Law

The Heaps graphs for unigrams, bigrams and trigrams of the diacritized and undiacritized texts are presented in Figures 2a, 2b and 2c. The Heaps curves of the diacritized and undiacritized texts presented on the three graphs drifted apart from the origin. However, the differences exhibited by the Heaps curves of diacrtized and undiacritized texts reduce as the n-grams increases while the graph becomes more linear. The Heaps exponent also increased as the n-grams increased with the undiacritized texts having higher exponential values. The Heaps exponents are expected to be close to 1, the trigrams have the highest exponents with 0.88 for the diacritized and undiacritized texts while the unigrams had the lowest exponents with 0.72 and 0.77 for the diacritized and undiacritized texts respectively. This shows that the trigrams exhibited the Heaps properties more than the bigrams and unigrams. This study shows that diacritics signifcantly affect word distribution in the Yoruba texts. This difference reduces as the co-occurred words (n-grams) under consideration increases.

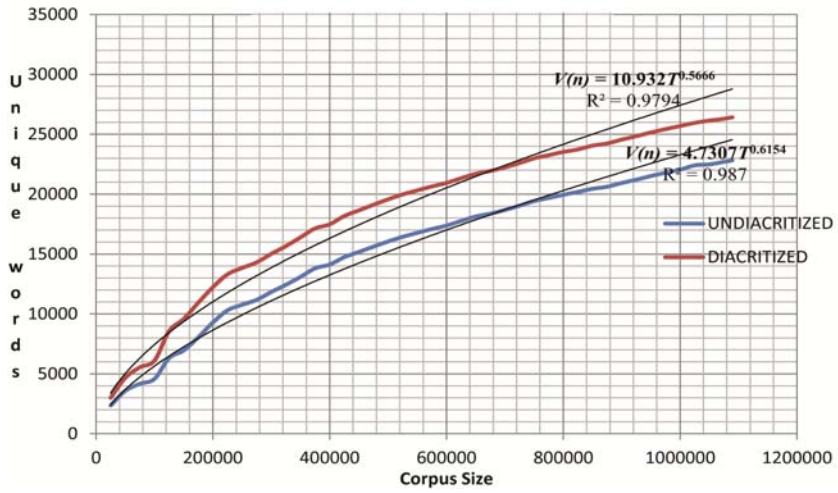

Figure 2a. Heaps' Graph for Unigrams

5. Conclusion and Recommendation

Zipf's and Heaps law are popular laws which are used in Natural Language Processing for modeling languages. It explains the characteristics of a language in relations to the increase in its vocabulary as the size of its texts increases. They present hidden natural regularities in statistical models. Heaps exponent for a language is a unique value which is language dependent and a distinguishing factor between languages. Hence, the behavior or the model of a language based the heaps law should portray the uniqueness of the language. In this case, the behavior of the language using the versions of Yoruba texts is a statistical account which suggest that the diacritics is a special feature which can affect the model or behavior of the language for language modeling. Though Zipf s model presented in this study present dissenting view as it does not explicate differences in the

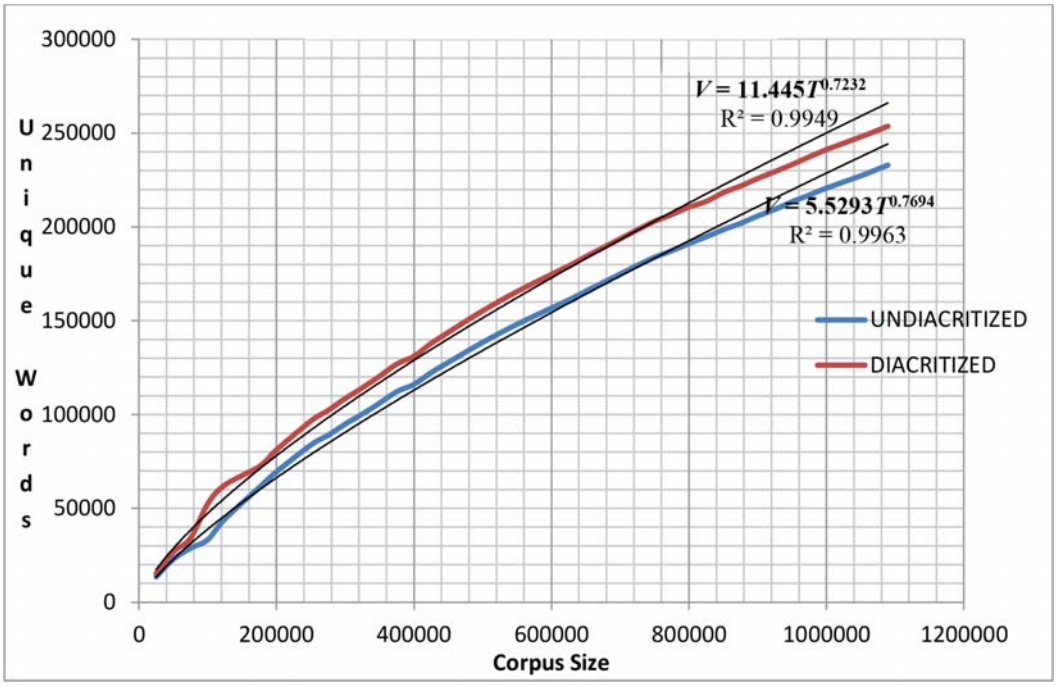

Figure 2b. Heaps' Graph for Bigrams

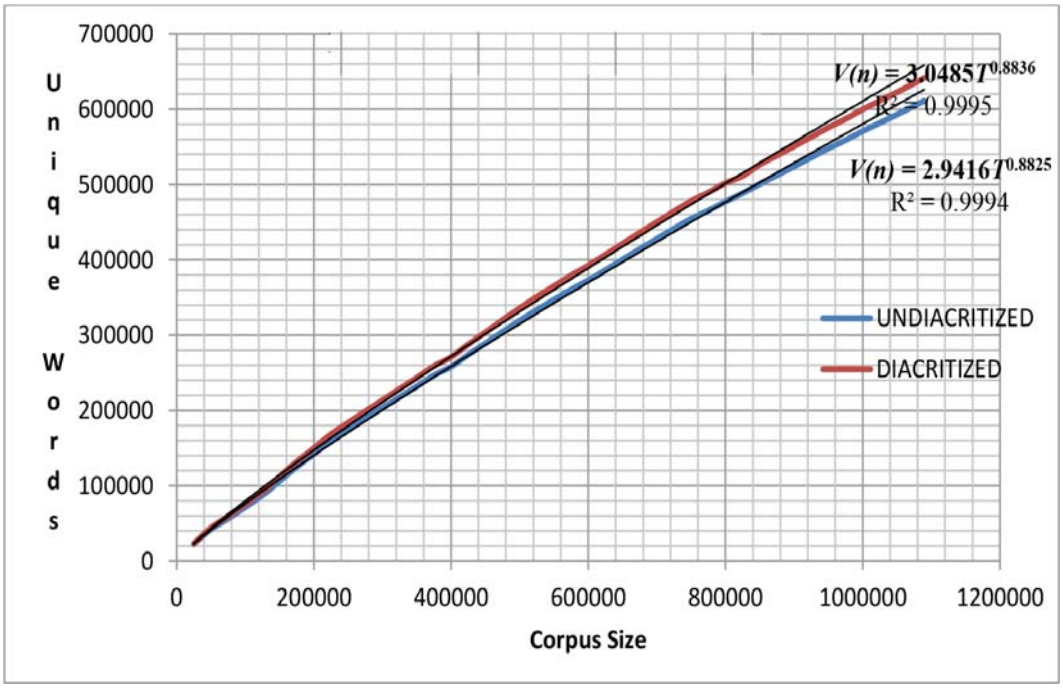

Figure 2c. Heaps' Graph for Trigrams

diacritized and undiacritized texts. As a suggested future research work, the explanations for the difference behaviours exhibited by the diacritized and undiacritized texts could be explored.

It further suggests that Yoruba language corporal for NLP studies are necessarily consistent in diacritics usage for accurate model of the language. A body of Yoruba texts that is partly diacritized will provide invalid (statistical) model and miscued behavior of the language and ultimately wrought wrong results for any NLP study. Furthermore, results of NLP studies on undiacritized version of the language texts cannot be extended to its diacritized version. For instance, [23] created stopword list

for both diacritized and undiacritized versions of the same corpus.

This research work proves that the Heap s law is dependent on the consistency of the use of orthography of a language. However, the dependency reduces as the number of n-grams increases while this effect was not exhibited on Zipfian s graph.

References

[1] Rosenfeld, R. (2000). Two decades of statistical language modeling: Where do we go from here? School of Computer Science Carnegie Mellon University, Pittsburgh, PA 15213, USA.

[2] Xu, P., Karakos, D., Khudanpur, S. (2009). Self-Supervised Discriminative Training of Statistical Language Models.

[3] Heaps, H. S. (1978). Information Retrieval: Computational and Theoretical Aspects. Orlando, FL, USA: Academic Press, Inc.

[4]Zipf, G. (1936). The Psychobiology of Language. London: Routledge.

[5] Zipf, G. (1949). Human behavior and the principle of least effort. Oxford, England: Addison-Wesley Press.

[6] Shamilov, Yolacan. (2006). Statistical Structure of Printed Turkish, English, German, French, Russian and Spanish, in Proceedings of the 9th WSEAS International Conference on Applied Mathematics, Istanbul, Turkey, 638–644.

[7] Géza, N., Csaba, Z. (2007). Multilingual Statistical Text Analysis, Zipf s Law and Hungarian Speech Generation. Department of Telecommunications & Telematics, Budapest University of Technology and Economics, Hungary.

[8]Manaris, B., Pellicoro, L., Pothering, G., Hodges, H. (2006). Investigating Esperanto s Statistical Proportions Relative to other Languages using Neural Networks and Zipf s Law," in Proceedings of the 2006 IASTED International Conference on ARTIFICIAL INTELLIGENCE AND APPLICATIONS (AIA 2006), February 13 – 16, 2006, Innsbruck, Austria.

[9] Damian, H., Marcelo, A. (2008). Dynamics of text generation with realistic Zipf distribution. Consejo Nacional de Investigaciones Cient´1ficas y T´ecnicas, Centro At´omico Bariloche and Instituto Balseiro, 8400 San Carlos de Bariloche, R´1o Negro, Argentina.

[10] Alexander, G., Grigori, S. (2001). Zipf Heaps and Laws' Coefficients Depend on Language, in Conference on Intelligent Text Processing and Computational Linguistics, February 18–24, 2001, Mexico City. Lecture Notes in Computer Science, Mexico City, 2001, 332–335.

[11] Bochkarev, V. V., Lerner, E. Y., Shevlyakova, A. V. (2014). Deviations in the Zipf and Heaps laws in natural languages, in Journal of Physics: Conference Series, , 490, 01.

[12] Wentian, L. (1992). Random Texts Exhibit Zipf s-Law-Like Word Frequency Distribution, *IEEE Trans. Inf. Theory*. 38 [6], p. 1842–1845.

[13]Asubiaro, T. (2011). An Analysis of the Structure of Index Terms for Yorùbá Texts, A Master s degree project, University of Ibadan, Africa Regional Centre for Information Science.

[14] Lu, L., Zhang, Z.K., Zhou, T. (2013). Deviation of Zipf's and Heaps Laws in Human Languages with Limited Dictionary Sizes, Sci. Rep., 3, 1–9.

[15] Xiao, H. (2008). On the Applicability of Zipf s Law in Chinese Word Frequency Distribution, *J. Chin. Lang. Comput.*18 [1], 33–46.

[16] Font-Clos, F., Boleda, G., lvaro Corral, A. (2013). A scaling law beyond Zipf's law and its relation to Heaps law, *J. Phys.*, 15.

[17] Van Leijenhorst, D. C., Van der Weide, T. P. (2005). A formal derivation of Heaps Law, Inf. Sci., 170, 263–272.

[18] Petersen, A. M., Tenenbaum, J. N., Havlin, S., Stanley, E., Perc, M. (2012). Languages cool as they expand: Allometric scaling and the decreasing need for new Words, Sci. Rep., 2.

[19] Eliazar, I. I., Cohen, M. H. (2012). Power-law connections: From Zipf to Heaps and beyond, *Ann. Phys.*, 332, p. 56–74.

[20] Eliazar, I. (2011). The growth statistics of Zipfian ensembles: Beyond Heaps law, *Phys. Stat. Mech. Its Appl.*, 390 [20], p. 3189–3203.

[21] Adesola,O. (2005). Yorùbá: A Grammar Sketch: Version 1.0. Rutgers University, U.S.A, 2005.

[22] Akilimali, F. (2008). Keyboard to help save Yorùbá and other endangered African languages.

[23] Asubiaro, T. (2013). Entropy-Based Generic Stopwords List for Yoruba Texts, *Int. J. Comput. Inf. Technol.* 2 [5], p. 1065–1068,

Biography

ASUBIARO, Toluwase works in the Systems Unit of E. Latunde Odeku Medical Library, College of Medicine, University of Ibadan, Nigeria as an Academic Librarian. His research interest is Information Retrieval, Statistical Language Modelling, Informetrics, Information systems and technology use. He had a B. Sc in Mathematics and a Masters' degree in Information Science.